第 40 卷 第 5 期 计 算 机 工 程 2014 年 5 月
Vol.40 No.5 Computer Engineering May 2014

·开发研究与工程应用· 文章编号：1000-3428(2014)05-0295-04 文献标识码：A 中图分类号：TP39

基于科研在线文档库平台的标签推荐系统

蔡 芳[1,2]，沈 一[1,2]，南 凯[1]

(1. 中国科学院计算机网络信息中心，北京 100190；2. 中国科学院大学，北京 100049)

摘 要：科研在线文档库是一个面向团队的文档协同与管理工具，为虚拟团队提供合作平台。它采用标签系统的方式组织其中的所有文档。在文档库的使用过程中，出现了无标签文档数量的累积以及用户为文档添加的标签质量偏低问题，影响文档的分类和共享。针对该问题，采用适用于科研在线文档库平台的标签推荐方法，包括协同过滤以及关键词抽取 2 个部分，促使用户为文档添加合格的标签，提高文档系统的使用效率。协同过滤推荐部分的实验采用准确率和召回率衡量标准，关键词抽取部分采用用户调查的实验方式，实验证明为每个文档提供 3 个候选标签能够得到理想效果。在实际使用环境中，该系统具有较高的精确度和可靠性，简单易于实现。

关键词：标签推荐；标签系统；协同过滤；关键词抽取；冷启动；文档协同

Tag Recommendation System Based on Duckling Document Library Platform

CAI Fang[1,2], SHEN Yi[1,2], NAN Kai[1]

(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】Duckling Document Library(DDL) is a tool for document collaboration and management among research teams. It provides a cooperation platform for virtual teams. Tag system is used to manage all the documents on it. During the use of the library, the number of documents without any tags is gradually accumulating and the quality of tags labeled by users to some documents is not so good. All these troubles impede the effective control of the documents. In order to solve these problems, this paper proposes a tag recommendation method suitable for the document library of research online platform, which includes collaboration filtering recommendation and keywords extraction recommendation, in this way users are prompted to add qualified tags and improve the efficiency of the document library. Precision and recall rate metrics are used in the collaboration filtering recommendation and user survey in the keywords extraction recommendation. Experimental results show that a recommended list of three tags can get desired effect. In production environment, this tag recommendation system has qualified accuracy, reliability and is easy to be implemented.

【Key words】tag recommendation; tag system; collaborative filtering; keywords extraction; cold-start; document collaboration

DOI: 10.3969/j.issn.1000-3428.2014.05.061

1 概述

Web2.0 下，用户行为由 Web1.0 中获取信息转变为以交互为主的方式，信息发布的来源转向 Web 用户。相对于传统的基于网站预先设定的分类体系的信息分类方法，标签系统的开放性、简单性、标签由资源共享者提供等特点[1]，使得它成为 Web2.0 网站的重要信息分类和索引方式。用户生成内容(User Generated Content, UGC)标签系统，通过让用户对信息打标签，将具有相同标签的信息进行分类归纳整理，形成以标签为中心的信息分类系统[2]。2004 年，标签系统领域的信息架构专家，提出分众分类法的概念，指群众自发性定义的平面非等级标签分类，用于信息的分类和共享。目前比较流行的 UGC 标签系统有书签类站点 Delicious、论文书签网站 CiteULike、相片分享网站 Flickr 等。

科研在线文档库(Duckling Document Library, DDL)是一个面向虚拟组织的协作式、文档共享和管理工具[3]。系统利用用户添加的标签对团队中所有的文档进行分类。其中未打标签的文档被放置于无标签文档类。一方面，随着团队成员和文档数量的增加，无标签文档的数量开始累积，这些文档处于一种平行无清晰组织结构的状态，当用户需要在其中寻找某一特定类别的信息时，比较耗时，这种情况不利于 DDL 文档的高效利用和管理，所以为无标签文档

基金项目：中国科学院十二五信息化基金资助项目“科研信息化应用推进工程(XXH12503)。

作者简介：蔡 芳(1990－)，女，硕士研究生，主研方向：网络协同，推荐系统；沈 一，博士研究生；南 凯，研究员。

收稿日期：2013-03-05 **修回日期：**2013-05-03 **E-mail：**caifangzky@sina.cn

推荐标签成为一种需求。另一方面，由于用户可以任意地为文档添加标签，而用户自身对信息和词汇的理解存在不准确性，使系统中的标签存在一定程度的冗余性、不一致性和不完备性[4]。这些问题都会影响到标签系统在进行文档组织、分类时的性能，所以提升标签的质量成为标签系统中核心的问题。当用户想为文档添加标签时，为用户提供高质量的标签备选，可以有效地缓解上述问题。

本文基于协同推荐的方式，为无标签页面提供高质量候选标签。如传统的协同推荐一样，对于一个新的团队文档集合，存在数据稀疏的冷启动问题。针对这种现象，系统采用关键词抽取的方式，利用文档自身的内容信息提取候选标签集合。当系统中的标签积累到一定质量和数量之后，再采取协同过滤的方式进行标签推荐。

本文利用文档内容信息和文档与标签之间的关系进行标签的推荐，而传统的标签推荐系统，基本都是基于用户、标签、资源 3 个对象之间的关系[5-6]，较少考虑资源自身的内容特征。当用户在 DDL 中对某一文档进行添加标签的操作时，系统会提供相关的推荐标签集合，此时，用户可以直接选择相关的标签进行添加，也可以在候选标签的提示下，添加自己的语义层面标签，这样可以有效地提升用户打标签的质量，降低打标签的难度。

2 标签推荐系统相关工作

标签推荐可以有效地提高系统标签质量，减少用户打标签的难度，近年来成为学术界和工业界关注研究的重点。在传统的标签推荐系统中，比较简单的标签推荐方法包括 4 种(统称为基于最流行的推荐法)：为用户推荐整个系统最热门的标签，为用户推荐他自己经常使用的标签，为用户推荐资源上最热门的标签。通过系数将前面 2 种方式的推荐结果进行线性加权的简单混合推荐[2]。

这 4 种方式不用进行复杂的模型训练和计算，实现成本低，在商业系统中较常使用。例如豆瓣，用户可以为一本书或者是一部电影添加标签，此时，标签系统会为用户提供 2 类标签，一类是用户自己的标签，另一类是此书籍或者电影上经常被标记的标签。对于商业产品，此类方法效果较好而且实现简单快速。但是这些算法对于新用户或者是不太热门的物品，存在冷启动问题，很难有较理想的推荐效果。

图模型也可以用于标签推荐系统。先根据用户对资源打标签这种行为，生成用户-资源-标签无向图。基于此图的相关算法有 FolkRank 算法[7]，此算法认为一个标签如果标记重要资源，而且是重要的用户进行的标注，那么这个标签就更重要。经过迭代计算，得到标签的得分排名，然后为资源提供 topN 标签推荐。另外一类是采用基于随机游走的 PersonalRank 算法[8]，此算法基本思路是：从用户 U 对应的节点 V_U 出发进行随机游走，游走到任何一个节点时，按照概率选择继续游走或者是返回节点 V_U 开始重新游走，经过迭代计算，使各个节点被访问的概率收敛到一个值，该概率就是推荐列表中标签的权重。这些算法都存在要进行模型训练、计算复杂、时间复杂度高等问题，在实际系统中应用起来还有很多实际的困难需要解决。

本文提出了一种综合协同过滤推荐以及关键词抽取的标签推荐方式。在 DDL 平台上，由于文档上被标记的标签都是共享的，即只存在文档、标签二维空间，而不是图模型中的三维空间，这样前文所说的一些推荐方式并不适合 DDL 实际环境，在此情况下本文提出一种不考虑用户的协同推荐方式，简单高效，易于实现。现在主流的标签推荐研究都是在 Delicious、Bibsonomy 等公开的数据之上进行的[9]，标签数据量有一定的基础，不用考虑冷启动的问题。在 DDL 中，若成立一个新的科研团队，其中基本没有标签，此时，采用第 2 种推荐方法：基于内容的关键词抽取标签推荐方法。

3 综合协同过滤和关键词抽取的标签推荐系统

Delicious、豆瓣等系统中，用户和资源之间是多对多的关系，用户 U1 和 U2 都可以对资源 I 添加标签，并且他们添加的标签集合 S1、S2 是独立的。而在 DDL 中，由于 DDL 的宗旨是团队协作和共享，团队成员之间的关系是十分亲密的，因此所有用户对于一个文档添加的标签都属于一个集合 S。由于不存在完整的用户-资源-标签三维空间，本文第 2 节中提到的主流标签推荐方式并不适合 DDL，从可用性、实用性、易于实现等方面考虑，提出一种综合协同过滤和关键词抽取的标签推荐方法。

当团队中已打标签的文档数目占所有文档的比例超过一个阈值时，采用协同过滤标签推荐方式，当小于这个阈值时，采用关键词抽取方式。

3.1 基于内容的协同过滤标签推荐

传统的协同过滤中，通过用户对资源的评分矩阵计算资源相似度或者是用户相似度。例如电子商务网站中当 2 个物品被同一个用户喜欢，那么它们的相似度加一。在 DDL 中，文档的协作分享面向科研团队，在一个团队中，用户和文档之间关系的黏度是比较强的，即一个用户访问某 2 个页面的可能性很大，并不能代表这 2 个页面的相似度关系，因此，使用传统的相似度判断方法并不适合 DDL。基于此，本文采用基于内容判断文档相似度的方法。

3.1.1 文档特征向量

对于 DDL 团队中的文档，在对其文档内容分词之后，利用 TF-IDF 模型计算文档中每个关键词的权重，然后构建文档特征向量：

$$\boldsymbol{D}_i=\left(\langle term_{i1},w_{i1}\rangle,\langle term_{i2},w_{i2}\rangle,\cdots,\langle term_{in},w_{in}\rangle\right)$$

其中，$\boldsymbol{D}_i$ 表示文档 i 的特征向量；$term_{ij}$ (j=1,2,⋯,n)表示将文档 i 的特征词按照权重由大到小排序之后的第 j 个特征词；w_{ij} 是其对应的 tf-idf 权重。

3.1.2　相似文档集合

目标是计算目标文档的相似文档集合。在构建了团队文档向量空间模型之后，利用余弦定理计算 2 个文档特征向量之间的距离：

$$sim(\boldsymbol{D}_i,\boldsymbol{D}_j)=\frac{\sum_{term_{im}=term_{jn}} w_{im}\times w_{jn}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2}\times\sqrt{\sum_{k=1}^{n} w_{jk}^2}}$$

其中，分子代表特征向量 $\boldsymbol{D}_i$ 和 $\boldsymbol{D}_j$ 中相同的特征词对应的权重乘积求和。

在 DDL 团队中，对于目标页面 d，计算它与团队中其他文档的相似度，选取前 30 个页面形成 d 的相似页面集合 N_d：

$$N_d=\{(\boldsymbol{D}_1,sim_{1d}),(\boldsymbol{D}_2,sim_{2d}),\cdots,(\boldsymbol{D}_{30},sim_{30d})\mid \boldsymbol{D}_i\in N_{\text{top30}},i=1,2,\cdots,30\}$$

其中，N_{top30} 表示与目标文档 d 相似度最大的前 30 个文档集合；$\boldsymbol{D}_i$ 表示第 i 个文档向量；sim_{id} 表示文档 i 与目标文档 d 的相似度权重。

3.1.3　推荐标签集合

在 DDL 中，对于目标文档 d，其相似文档集合为 N_d，对于其中的每个文档 i，其上有一些已经被标记上的标签 t，将对应于 i 的已有标签集合记为 T_i。对页面 d 的推荐标签集合如下：$T_{rec-d}=\{(t_{d1},wt_{d1}),(t_{d2},wt_{d2}),\cdots,(t_{dk},wt_{dk})\}$。其中，$t_{di}\in T_1\cup T_2\cup\cdots\cup T_{30}$ (i=1,2,⋯,k)是为目标文档 d 推荐的第 i 个标签；wt_{di} 是标签 t_{di} 对应的排名权重，由如下公式计算：

$$N_{t_{di}}=\{(\boldsymbol{D}_k,sim_{kd})\mid(\boldsymbol{D}_k,sim_{kd})\in N_d,t_{di}\in T_k\}$$

$$wt_{di}=\sum_{(\boldsymbol{D}_k,sim_{kd})\in N_{t_{di}}} sim_{kd}$$

其中，T_k 代表文档 k 上已有的标签集合；$N_{t_{di}}$ 代表在目标文档 d 的相似文档集合 N_d 中包含标签 t_{di} 的所有文档的集合；$T_{rec\text{-}d}$ 按照标签权重 wt_{di} 进行排序。

3.2　关键词抽取

用 TF-IDF 度量关键词的权重。采用公式 $tf-idf_{t,d}=tf_{t,d}\times idf_t$，$tf_{t,d}$ 表示词项频率，idf_t 表示逆文档频率。在词袋模型[10]的文档视图下，TF-IDF 模型能够表示文档中词项的区分度和重要度[11]。TF-IDF 被公认为信息检索中最重要的发明，常用于搜索引擎排名中确定网页和查询的相关性、自底向上文档分类等问题中[12]。

对于一个全新的团队，系统中基本没有标签，在协同过滤方式中会出现冷启动的问题，本文采用关键词抽取的方式来解决。具体做法如下：采用 IKAnalyzer 中文分词器的智能切分方式对文档分词，将 DDL 中已经存在的标签作为自定义的扩展词典，过滤单个汉字词项和数字，然后统计文档中词项的 TF-IDF 值，选取 topK 作为推荐集合：

$$S_{rec}=\{t_1,t_2,\cdots,t_k\}$$

其中，关键词按照权重由大到小排名，推荐文档的前 K 个最大 TF-IDF 权重的关键词集合。

4　实验及结果分析

4.1　基于内容的协同过滤标签推荐

4.1.1　实验数据及度量方法

为验证算法的性能，本系统采用 DDL 中某一团队的部分数据集合。这个数据集合包含 3 000 个页面。随机选取所有页面的 20%作为测试集合，即训练集合页面数目为 600。

由于系统属于 TopN 推荐，即为用户提供一个推荐列表。TopN 推荐的预测精度一般通过准确率和召回率来度量。这里，采用这 2 种传统的度量方式：

$$Precision=\frac{\sum_{p\in P}|R(p)\cap T(p)|}{\sum_{p\in P}|R(p)|}$$

$$Recall=\frac{\sum_{p\in P}|R(p)\cap T(p)|}{\sum_{p\in P}|T(p)|}$$

其中，p 表示测试页面集合；$R(p)$表示给页面推荐的标签集合；$T(p)$表示测试集中的页面实际被标记上的标签。

通过选取不同的列表长度 N，计算出一组准确率和召回率，以此判断最佳的推荐长度。为了保证测试实验的准确性，重复实验 5 次，每次用于测试的 600 个页面都是随机选择的不同页面。

4.1.2　结果分析

选取 N={3,4,5,6}进行实验，每次进行 5 次重复实验。图 1 代表取不同的 N 值时的准确率，图 2 是对应的召回率。

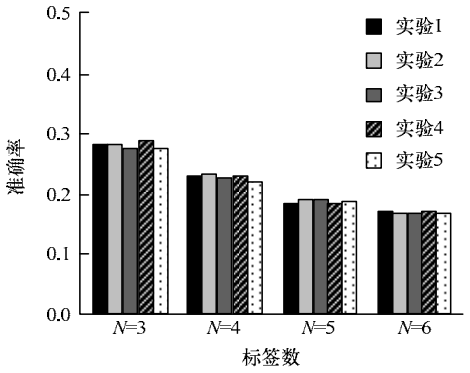

图 1　N 取不同值时的准确率

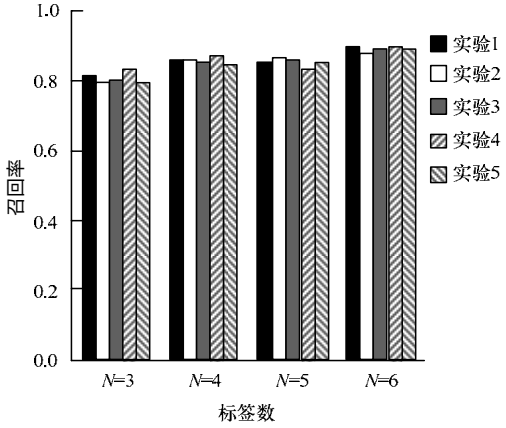

图 2　N 取不同值时的召回率

从图 1 中可以看出，准确率相对于召回率处于一个较小的取值空间，因为准确率代表的是页面推荐集合和原有标签集合的交集 C 与推荐标签集合的总数目 R 的比例。当 N 变大时，R 增长较快，例如 N 为 3 时，推荐总数为 3×600=1 800，N 为 4 时推荐总数为 4×600=2 400，而选用的团队页面集合基本上每个页面的标签数目在 1 个~2 个之间，而集合 C 受到原有标签集合的数目限制，C 与 R 的数量差距较大，这也就解释了精确率都在较小数据区间内的现象。

而准确率随着 N 的增长呈现下降的趋势，主要是由于 N 的增长导致 R 显著增大，但是对于 C 的提升没有很明显的效果，出于实际 DDL 中页面的标签基本上是在 3 个以内，此处认为选择 N 为 3 时，比较理想。

召回率代表了集合 C 与页面原有标签集合 T 的比例。对于随机选择的 600 个测试页面集合，T 的数量基本稳定，但是当增大推荐数目 N 时，如同在分析精确度时所描述的，N 对于推荐效果的提升虽然没有很显著的影响，但是当推荐的候选集合增大，交集 C 还是会有小幅度的增加，因此，也就表现为召回率的小幅度增大变化，但是这个增长幅度太小，故认为 N 为 3 时的召回率已经是比较理想了。

综合上述原因，采用推荐标签集合长度 N 为 3 较理想。

4.2 关键词抽取推荐

本文是基于 TF-IDF 进行关键词提取，所得到的关键词推荐集合是基于分词结果。例如页面“试用期/实习期管理”，得到的推荐集合是{实习期，试用期，转正}；页面“考勤公示说明”，推荐集合{缺勤，考勤，公示}；页面“2010 级硕士生开题答辩”，推荐集合{开题，硕士生，2010 级}。可以看出，内容抽取的方式能够得到一些比较好的代表文档内容的关键词，这样能够方便用户对文档添加标签。内容抽取方式得到的是词粒度的标签。而当 DDL 团队被使用一段时间之后，部分页面会被添加上一些语义层面的标签，例如“科研与教育”、“全室共享”，这样在基于内容的协同推荐方式下，就会为页面提供一些语义层面的标签推荐，例如上面提到的页面“2010 级硕士生开题答辩”，得到推荐集合{科研与教育，分享与研究，默认集合 }。

对于该推荐方式采用用户调查的方式进行实验。由于对于已有标签的页面，其上的标签可能会影响用户对推荐结果的主观判断。因此，随机选择团队中个 300 个未打标签页面，选择 5 个用户参加调查，评价分为 3 个等级。重复实验 5 次结果如表 1 所示。其中数据分别代表 300 个页面中用户满意、感觉一般和不满意页面的数目所占的比例。

表 1 用户调查满意度

实验	页面数	满意/%	一般/%	不满意/%
1	300	65	27	8
2	300	60	28	12
3	300	64	27	9
4	300	60	30	10
5	300	63	30	7

随着使用时间的增长，标签数量和质量会逐步的积累和改善，从而标签推荐系统的效果也会稳步上升。

5 结束语

本文综合协同过滤方法和关键词抽取方法对 DDL 团队文档推荐标签。在解决标签推荐冷启动问题的同时能够为用户提供高质量的候选标签集合，方便用户对页面添加具有代表性的标签。提升了 DDL 的标签系统，使得文档的组织、管理和分享更加高效有序。实验结果证明，该系统能够为文档提供较高精度的标签推荐，有利于 DDL 标签系统的有效构建和发展。下一步工作着重于提高标签推荐的精度，同时在关键词抽取方面，利用主题模型进行实验，和 TF-IDF 方法进行对比。

参考文献

[1] Golder S A, Huberman B A. The Structure of Collaborative Tagging System[J]. Journal of Information Science, 2006, 32(2): 198-208.

[2] 项 亮. 推荐系统实践[M]. 北京: 人民邮电出版社, 2012.

[3] 南 凯, 董科军, 谢建军, 等. 面向云服务的科研协同平台研究[J]. 华中科技大学学报: 自然科学版, 2010, 38(1): 14-19.

[4] Guy M, Tonkin E. Folksonomies: Tidying up Tags?[J]. D-Lib Magazine, 2006, 12(1): 1-15.

[5] 许棣华, 王志坚, 林巧民, 等. 一种基于偏好的个性化标签推荐系统[J]. 计算机应用研究, 2011, 28(7): 2573-2575.

[6] Gemmell J, Schimoler T, Mobasher B, et al. Hybrid Tag Recommendation for Social Annotation Systems[C]//Proc. of the 19th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2010: 829-838.

[7] Hotho A, Jäschke R, Schmitz C, et al. Information Retrieval in Folksonomies: Search and Ranking[C]//Proc. of the 3rd European Semantic Web Conference. Berlin, Germany: Springer-Verlag, 2006: 411-426.

[8] Haveliwala T H. Topic-sensitive PageRank[C]//Proc. of the 11th International Conference on World Wide Web. New York, USA: ACM Press, 2002: 517-526.

[9] 勒延安, 李玉华, 刘行军. 不同粒度标签推荐算法的比较研究[J]. 计算机应用研究, 2012, 19(2): 504-509.

[10] Lewis D D. Naive(Bayes) at Forty: The Independence Assumption in Information Retrieval[C]//Proc. of the 10th European Conference on Machine Learning. London, UK: Springer-Verlag, 1998: 4-15.

[11] Manning C D, Raghavan P, Schütze H. 信息检索导论[M]. 王 斌, 译. 北京: 人民邮电出版社, 2010.

[12] 吴 军. 数学之美[M]. 北京: 人民邮电出版社, 2012.

编辑 顾逸斐